

**PAKDD COMPETITION 2007**

# **Predictive Modeling Cross Selling of Home Loans to Credit Card Customers**

Hualin Wang<sup>1</sup>

Amy Yu<sup>1</sup>

Kaixia Zhang<sup>1</sup>



800 Tech Center Drive  
Gahanna, Ohio 43230, USA

*April 11, 2007*

## ***Outline of Approach***

Our approach employs PROBIT regression modeling and ensembles a set of PROBIT models. All the analyses were performed using SAS<sup>®</sup> 2. We perform a large amount of univariate and bivariate analyses for handling missing values, capping extreme values, binning, variable transformations and creating interactions.

A natural candidate for modeling the cross-selling propensity is the class of PROBIT regression models:

$$p = \text{probability}(Y = 0) = C + (1 - C) * F(X' \beta)$$

where

$p$  : the probability for the response to be 0;

$C$  : the natural response rate for the PROBIT model;

$X$  : a set of explanatory variables;

$\beta$  : a vector of parameter estimates; and

$F$  : a link function, usually a cumulative distribution function (e.g., the normal, logistic function or extreme value)

Our comparative study shows that using a cumulative function of the normal distribution tends to have higher and more stable c-statistics for different random samples in this case. Other analysis helps to determine the range of weights for the PROBIT model. The final set of weights used is the 10 integers from 3 through 12.

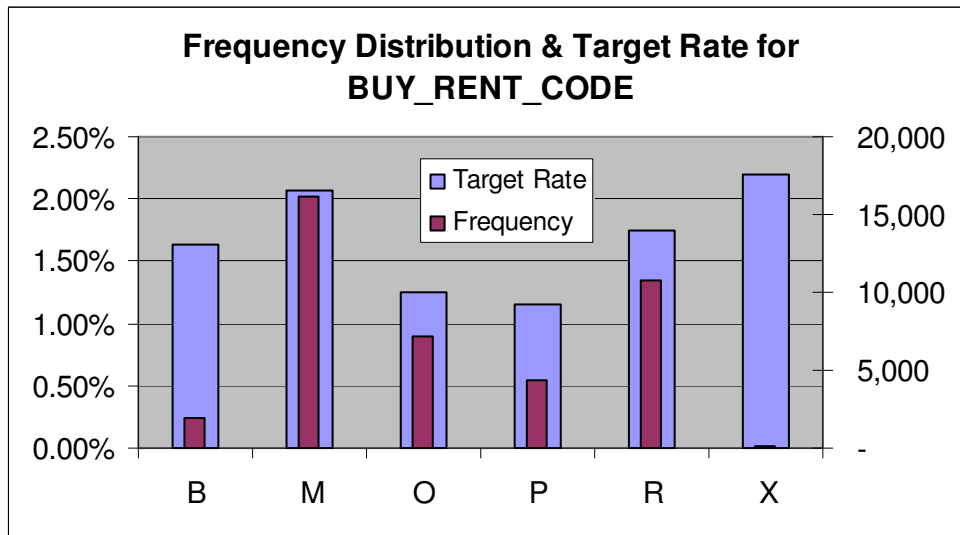
With these factors, our final model is built by following these 2 steps:

1. Pick any integer for weight between 3 and 12, build an ensemble of 10 PROBIT models using 10 bootstrapped samples and average the 10 probabilities. This is the model for the selected weight. At the end of this process, there are 10 ensemble models corresponding to the 10 different weights (from 3 through 12).
2. For each observation, remove the largest as well as the smallest probabilities and compute the mean probability of the remaining 8 probabilities. This average probability based on a scoring mechanism similar to a diving scoring system is the final predicted value.

## Exploratory Data Analysis

There are 40 possible raw explanatory variables, but B\_DEF\_UNPD\_L12M is equal to 0 for all records. Since there are not many raw attributes, one of the challenge particularly for this problem is to create more predictive variables out of the 39 raw variables.

For categorical variables, check their frequency distributions and the target rate which is the percent of records with TARGET\_FLAG = 1. Based on the results, we do some *regrouping* or *binning*. For example, the Frequency Distribution & Target Rate for BUY\_RENT\_CODE is shown below:



The 'X' valued category is extremely small (only 137 records out of 40,700). It should be recoded as one of the other values. In this case, we recode it as 'M'.

For numerical valued variables, we do *capping*, *ranking*, *Box-Cox transformations*, and other nonlinear transformations. For example, the ranks for CURR\_EMPL\_MTHS are shown here:

Rank	Frequency	Minimum CURR_EMPL_MTHS	Maximum CURR_EMPL_MTHS
0	3770	0	5
1	4710	6	12
2	3354	13	23
3	4406	24	35
4	3787	36	47
5	4704	48	60
6	3799	61	84
7	4251	85	120
8	3839	121	191
9	4080	192	1000

Based on this result, we create a new variable N\_CURR\_EMPL\_MTHS as follows:

```
IF CURR_EMPL_MTHS<=5 THEN N_CURR_EMPL_MTHS =0;  
ELSE IF CURR_EMPL_MTHS<=12 THEN N_CURR_EMPL_MTHS =1;
```

and so on.

Capping is also used to limit some relatively large values, for example:

```
N_NBR_OF_DEPENDANTS = Minimum (NBR_OF_DEPENDANTS, 5);
```

For numerical variables, Box-Cox transformations, especially, logarithm and square root are employed.

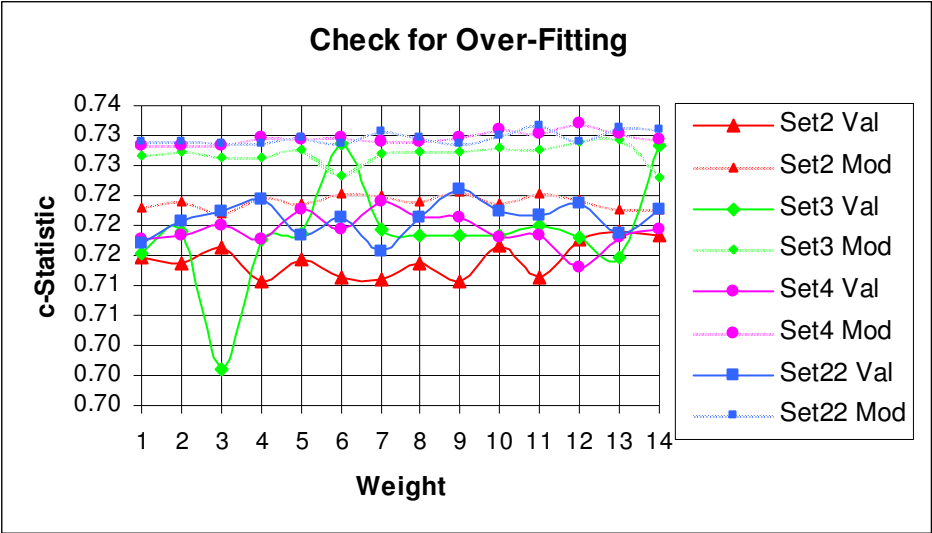
We tried to identify any first-order *interactions* and created some hypothesized interaction terms, but none is found to be significant. For categorical variables, we also sort the categories based on the target rates, and then created numerical variables.

## ***Variable & Model Selection and the Final Model***

At the end of the EDA, we created a few hundreds of variables. SAS procedures LOGISTIC and STEPDISC as well as Decision Trees in E-Miner are the main tools used for variable selection. We employed stepwise, backward, and best subsets of variables for selecting variables. We also gradually expand trees as a way for variable selection.

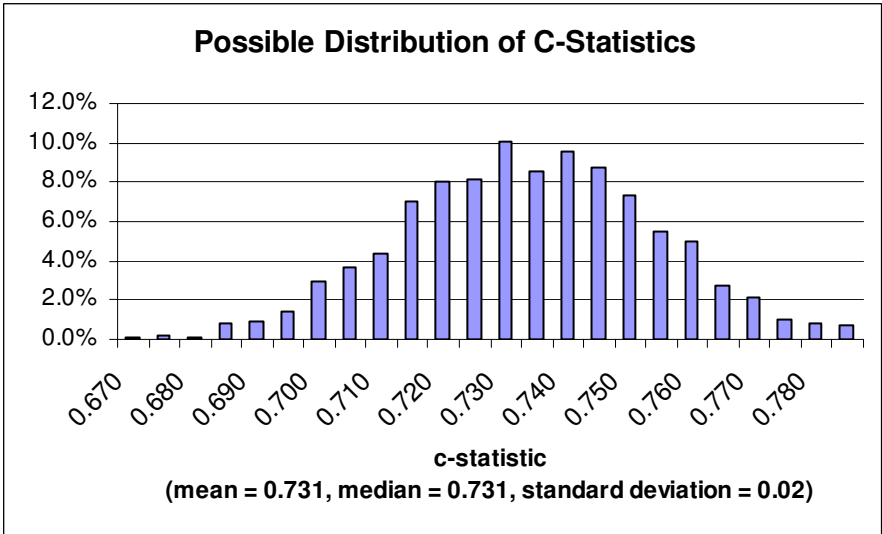
One of the major challenges we have is to build a stable model in terms of c-statistic on a sample of 8,000 records. To achieve this, we have to determine the number of variables as well as what variables in a particular model. We start with a logistic regression model. Take an 80% for modeling and 20% for validation. For every random data split, compare the c-statistics on modeling and validation datasets, and check the parameter estimates. Run through this comparison for multiple times and the parameter estimates are recorded. The mean, standard deviation and the coefficient of variation of each parameter estimate is computed. Their values help to determine which variables are to be removed if we think that there are too many variables in the model.

Another way employed for checking over-fitting is to compare the c-statistics on modeling and validation datasets. We gradually increase the number of variables in the model and compare the amount of the drop in c-statistics from modeling to validation datasets. An example of these charts is shown below. Four models are under consideration corresponding to four sets of variables which have different numbers of variables. It can be seen that the c-statistics on modeling datasets are very close, and it is the values on validation datasets that determine the number of variables for the model. We test gradually expanded sets of variables, and we find that when the number of variables in the model is about 16, which has about 1.6% drop in c-statistic from a modeling dataset to a validation one, the over-fitting is not serious and the model can achieve largest c-statistics on validations. In short, the key idea is to compare the changes in c-statistic on modeling datasets with those on validation datasets. As the number of variables in a model increases, the c-statistics on validation increase diminishingly or even start to decrease.



After comparing the PROBIT models using LOGISTIC function with those using the cumulative function of the NORMAL distribution as the link function, we decide to use the NORMAL. For the weight in the model, we compare a range of values and decide to use all integers from 3 through 12. In the PROBIT model specification,  $C$  is set to be 0. To build a final model, pick an integer and build an ensemble model by averaging the probabilities produced by 10 models where each of the 10 models is developed on a bootstrap sample of the 40,700 records. This process results in 10 ensemble models corresponding to the 10 weights. To predict, score each record using the 10 models. Then remove both the largest and the smallest value, and average the remaining 8 probabilities. This average is the final prediction of probability for the target to be 1.

We test the final model on 2000 random samples of 8,000 records. The c-statistics on the samples are distributed as follows:



This shows that the c-statistic on any particular sample of 8,000 records is in a wide range.

## **Conclusion**

Overall the target rate is about 1.72%. Applicability of the model depends on what the company would do and the profit margin for each action they would take for those customers identified by the model to have relatively higher probabilities for cross selling the home loans. Look at the following gains chart. It is created by ranking order the predicted probabilities and puts all records into 10 equal sized deciles. Decile 1, for example, has 246 home loan buyers and 3,824 non-buyers. The target rate is 6.04% which is 3.51 times 1.72%, the overall rate. The chart shows that the top three deciles all have higher gains than 1. These top three deciles can capture 63% of total 700 buyers.

Decile	Target Flag=1		Target Flag=0		Gain	
	By Decile	Cumulative %	By Decile	Cumulative %	By Decile	Cumulative
1	246	35%	3,824	10%	3.51	3.51
2	116	52%	3,954	19%	1.66	2.59
3	80	63%	3,990	29%	1.14	2.10
4	55	71%	4,015	39%	0.79	1.78
5	63	80%	4,007	49%	0.90	1.60
6	43	86%	4,027	60%	0.61	1.44
7	36	91%	4,034	70%	0.51	1.30
8	20	94%	4,049	80%	0.29	1.18
9	26	98%	4,045	90%	0.37	1.09
10	15	100%	4,055	100%	0.21	1.00
Total	700		40,000			

Some other facts may be worth a mention. The target rate is the highest at around the age of 30, and a bivariate also shows that, in general, the higher income tends to have a higher target rate. However, in general, income tends to be higher as age increases (to certain point). These two facts may show that the type of home loans the company is offering may meet some unique needs of some customers. If the loans are created to meet more general demands, we recommend the company to conduct a survey based research such as a conjoint analysis study to understand their customers' need. They can also start with segmenting customers to reveal the differences among their customers and then targeting more specifically. We strongly believe in 'Know More. Sell More'. In this case, the company needs to know more about their customers, to know what they need, and to gather more information from all possible sources. Ultimately the company will sell more, likely a lot of more.

## **Notes**

1. Hualin Wang, Amy Yu and Kaixia Zhang are in the Advanced Analytics group within Retail Services, one of the major businesses of Alliance Data ([www.alliancedata.com](http://www.alliancedata.com)). Hualin Wang is a Senior Statistician, Amy Yu a Senior Statistician, and Kaixia Zhang a Marketing Manager. They work for more than 100 clients whose businesses include, but are not limited to, specialty retail & department stores, healthcare, furniture, home improvement, and jewelers. Their work includes private label and co-brand credit card acquisition, portfolio management, marketing campaign design and analysis, customer segmentation, retention and re-activation.
2. SAS<sup>®</sup> is a registered trademark of SAS Institute Inc. in USA.